International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0470
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3181-3185

Study on AI Generated Fake-Media Detection

Vishal Gawali [1], Chaturdhan Chaubey[2], Mahesh Gaikwad[3], Akash Gidde[4], Nilesh Bhelkar[5]

[1,2,3,4]UG Scholar, Dept. of AI&DS, Rajiv Gandhi Institute of Tech., Mumbai, Maharashtra, India.

[5]Assistant Professor, Dept. of AI&DS, Rajiv Gandhi Institute of Tech., Mumbai, Maharashtra, India.

Emails: *vishalgawali5460@gmail.com[1] , cchaturdhan82@gmail.com[2], maheshgaikwad7678@gmail.com[3], akashgidde5800@gmail.com[4] , nilesh.bhelkar@mctrgit.ac.in[5]*

Abstract

The rapid growth of AI-generated images, especially with techniques such as Generative Adversarial Networks (GANs), has complicated the ability to tell apart genuine content from artificial creations. This issue is vital for preserving the authenticity of visual media, where conventional detection methods often struggle. Current detection approaches concentrate on machine learning and deep learning techniques, including neural networks (CNNs). These methods aim to reveal subtle flaws and irregularities in images, like inconsistencies in pixel distribution and lighting, which serve as critical signs of AI involvement. The research emphasizes the necessity for ongoing development in detection technologies to keep up with the quick progress of AI advancements. Ensuring that these detection methods are accurate and dependable is crucial for protecting against misinformation and maintaining confidence in digital content. This paper reviewed Deepfake detection system.

Keywords: *Deepfake, DeepfakeStack, GANs, Deep Ensemble Learning, Machine learning.*

1. Introduction

The swift development of artificial intelligence has led to some amazing things, but it's also brought up big problems. One of those is the rise of AI-made media, like deepfakes and synthetic pictures. These super-realistic fake images and videos are so good that telling what's real from what isn't is getting really tough. This is a big worry. It could help spread lies, fool people, and sway opinions in serious ways. At the heart of these technologies are what's called Convolutional Neural Networks (CNNs). They're super important for recognizing images. CNN models have changed over time. They started with simpler designs like LeNet5, and now we have more advanced ones like ResNets & DenseNets [1]. These newer models fix issues like vanishing gradients & overfitting by improving how information flows. But with all this progress in making realistic content, it's also made deepfakes more common. Deepfakes are fake audio and video that look real because of deep learning tricks [5]. Basically, AI looks at many pictures or videos of a person's face to swap that face onto someone else's body in an image or video. The result? Very convincing but totally fake stuff! Recently, two popular techniques for changing faces have caught a lot of attention—especially from people up to no good—raising worries about how this tech might be misused. To tackle these risks, researchers are working hard on smart detection systems to spot AI-made media. They use machine learning and deep learning to find things that don't seem quite right, helping separate real content from fake stuff. There's a real need for strong solutions to fight against how AI-generated media can be misused as it becomes more common every day.

2. Literature Survey

The research [1] has brought tremendous improvements in the digital face manipulation detection handling of face forgery as an area of study still has a long way to go. Research from 2019 showed modeling these unreliable head poses in

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0470
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3181-3185

deepfakes can enable models to obtain high AUROC scores, suggesting that deepfakes are indeed detectable with near perfect classification accuracy. These models in particular focused on the relative difference from central-face estimated head poses with respect to full-head and were marked as key discriminators for modified images. Nonetheless, the study highlighted major drawbacks with current methods in dealing especially those to low-quality images which most of the current techniques cannot tackle properly. The author in [2] is strengthening the notion of adaptability and specificity, in 2020 added an attention mechanism to their detection strategy. This new approach is more suitable for a problem such as remember where it allows to respond with enhanced detection and localization of manipulated facial features at the same time, particularly so in scenarios with decreased false detection rates. Yet, it also highlighted the need for broader datasets that cover a wider sweep of manipulation types to more comprehensively evaluate and improve detection methods. The researcher in [3] point to a way forward by improving both sensitivity and robustness of detection schemes against the intricate and diverse faces-in-the-wild manipulations in the digital era and at the same time are calling for novel methods from the research community to fill the gaps previously identified. Much of the research has highlighted the difficulty faced in detecting deepfakes and trying to reduce face manipulations by acknowledging image artifacts. The approach which they used was based on machine learning methods, i.e., classification algorithms k-NN and logistic regression for the analysis & segmentation of images that were identified as deepfake. When tested with GAN generated data, the k-NN classifier performed strongly (AUC: 0.852), highlighting its ability to discriminate between real and fake images. In addition, logistic regression models in combination with other features achieve comparable performance comparable to deep models, showing potential for simple and light deepfake detection the open-source software he used won the end-to-end experiments for face examination. This feature will just record the time point when a malicious ad appears. not guarantee to store names of people who have visited any site on daytimes face current online attacks. But even so, there obviously will need to be technological filings for quick recovery. The need for better adjustment of contact points on between structures and organs has often been raised in past experience with interventional systems. A general trend is less movable joints in medical equipment the researchers put forward the direction of post-operative drainage for patients who undergo hepatectomy. Many works of this kind are published these days but there is no cross-reference standard cataloging formal structure to account for them all so it's tough sledding indeed where these do not appear for discussion. In [4] Dense Net – a new and exciting concept in the construction of convolutional networks. DenseNets are known for their dense connectivity pattern within layers, which leads to a decrease in the number of model parameters compared to conventional architectures such as ResNets. This design not only solves the vanishing gradient problem through improving the gradient flow, but also improves the parameter efficiency, making DenseNets have high computational efficiency. For example, a DenseNet that has the same computational complexity as a ResNet-50 can outperform a ResNet-101, providing high computational efficiency. The authors pointed out several questions that could be asked in future works, including the exploration of the capacity of DenseNets that had not been systematically studied at the time of writing. The work in research [5] introduces DeepfakeStack, a new ensemble learning technique dedicated to improving the deepfake detection. This is where DeepfakeStack is superior because it applies several deep learning models in one step to enhance the detection rates. Thus, this ensemble approach takes the best from different models to design a reliable system that can detect deepfake videos with a high precision as confirmed by the detection accuracy of 99.65%. This not only provides an effective approach for deepfake detection but also presents possible directions for model enhancement and future deep learning research on combating fake media. The table 1 shows the details of literature survey.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0470
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3181-3185

Table 1 Literature Survey Detail

Author & Ref.	Year	Methods
Xin Yang, Yuezun Li and Siwei Lyu [2]	2019	Exposing Deepfakes Using Inconsistent Head Poses
Hao Dan FengLiu Joel Stehouwer Xiaoming Liu Anil Jain [5]	2020	On the Detection of Digital Face Manipulation
Falko Matern Christian Riess Marc Stamminger [1]	2019	Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations.
Laurens van der Maaten, Zhuang Liu, Gao Huang [3]	2017	Densely Connected Convolutional Networks
Md. Shohel Rana, Andrew H. Sung [4]	2020	DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection

2.1 Methods of Detecting Media

Head Pose Estimation: Deepfakes may fail to reproduce natural head movements and poses because they are not perfect. Inconsistencies in the head pose can therefore be useful in identifying manipulated media. [1]

Facial Landmark Analysis: This approach involves comparing and contrasting some facial characteristics (eyeballs, nose and mouth) in order to assess the common abnormalities and defects that are apparent in manipulated images. This is because the generated faces do not always preserve the proper positioning of these landmarks during expressions or head movements; hence, their movements can be used to identify manipulations. [2]

Visual Artifact Detection: This approach aims to detect digital imprints which are an outcome of the deepfake generation including improper illumination, anomalous texture or pixel level discrepancies. Such minor artifacts that can easily escape the notice of the human eye can be utilized by deep learning algorithms to determine if a given face image or video has been manipulated or not. [3]

DenseNet for Feature Propagation: Enhancing the feature spread in DenseNet is accomplished by ensuring that all layers are interconnected in a feed forward manner through densely connected convolutional layers. This architecture allows better gradient flow and reuse of features from earlier layers, making it highly effective for tasks like image classification, where detailed feature extraction is essential for detecting subtle visual differences. [4]

Deep Ensemble-Based Learning: A deep ensemble-based learning technique combines multiple deep learning models, each trained to detect different aspects of deepfakes (such as pixel-level anomalies, temporal inconsistencies, or facial distortions). [5-7]

3. Proposed System

Overview of the Proposed system: -

3.1 DeepfakeStack Technique

A method for deep learning to detect deep fake images. Fuses few states of the art deep learning classifiers and then uses it in a single classifier for better classification results [8].

3.2 Architecture

- The base learner models which were used are XceptionNet, ResNet101, InceptionResNetV2 and so on.
- The first level model selected is Deepfake Classifier (DFC) which will learn in the presence of second level base learner's prediction.

3.3 Model Training

- The individual predictions are then presented to the meta learner in order to induce knowledge from them [9-11].
- The meta-learners are trained with out of sample data, this is data that was not used when building the model.

4. System Architecture

Data Collection: We are downloading the data from the CelebFaces Attributes (CelebA) Dataset from Kaggle platform (Figure 1).

Data Preprocessing: It can be used in order to prepare the collected data and adjust it for processing in a more suitable form. This comprises imputation of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0470
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3181-3185

the missing values, scaling of features, coding of categorical data and dividing the data into training and test set [12-14].

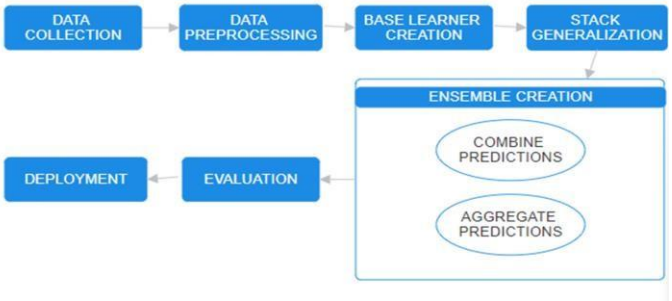


Figure 1 System Architecture

Base Learner Creation: Build multiple models (Machine Learning models) for the same problem using other different algorithms or different techniques.

Stack Generalization: A method where the result of the base learners is given as input to another learner in order to enhance the result [15].

Ensemble Creation: Standalone, simple, or more sophisticated methods can be employed to aggregate the forecasts coming from individual learners. It can also encompass the means by which the meta-learner obtains the results of the predictions made.

Evaluation: At least one of the parameters such as Accuracy, Precision, Recall or F1 Score recommended should be used to measure the efficiency of the proposed ensemble model. Check your results and rule out the likelihood of chance.

Deployment: After any model has been developed and tested, it ought to be planted into a live environment where it can generate its prediction on new data. This will entail developing the API or embedding the model into websites or programs.

5. Results and Conclusion

5.1 Results

There are many works dedicated to enhancing the techniques for identifying manipulations in images and videos and, in particular, Deepfakes, using classical and deep learning-based approaches. The author also tested simple visual artifacts that include eye color inconsistencies, lighting effects, and geometric deformities that were also found to be useful with an AUC of 0.866 [3]. A new attention-based mechanism enhanced the detection by learning the manipulated regions, achieving an AUC of 99.76% on the DFFD dataset [2]. Head pose discrepancies were also quite impactful with AUROC results of 0.89 for per frame and 0.974 for video level in the UADFV dataset and 0.843 in DARPA GAN Challenge dataset [1]. DeepfakeStack, an ensemble model that incorporates models such as XceptionNet and DenseNet, obtained an accuracy of 99.65% and an AUROC of 1.0 pointing out the model's capability [5]. The DenseNet's structure also reveal efficiency in its performance

Conclusion

The combined research suggests that basic visual cues, as well as complex deep learning algorithms, are quite efficient in identifying digital face manipulations such as Deepfakes. Techniques like head pose discrepancy, Attention based architectures, and ensemble of deep models have been seen to perform well with satisfactory generalization across the different datasets [16]. This is where architectures like DenseNet come into play; it has been shown that it is indeed possible to attain the same levels of accuracy as these larger and more complex counterparts while employing These studies indicate that, although existing detection methods are efficient, further developments are needed.

Acknowledgements

First and foremost, our appreciation to the AI media tools that have supported this work in its preparation. The potential of these technologies to generate, transform and analyze visual and audio materials helped to expand the understanding of the detection methods presented in this research. We are especially thankful for the advancements made in the field of Artificial Intelligence and deep learning methodologies that allowed for a better understanding of the manipulative nature of media, and thus created reliable methods for detecting such media. Additionally, we would like to thank the researchers and developers of open-source AI media generation platforms and datasets, which were used in this study.

References

[1] F. Matern, C. Riess, and M. Stamminger, (2019) ‘‘Exploiting visual artifacts to expose deepfakes and face manipulations,’’ in Proc.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0470
e ISSN: 2584-2854
Volume: 02
Issue: 10 October 2024
Page No: 3181-3185

IEEE Winter Appl. Comput. Vis. Workshops (WACVW), Waikoloa Village, HI, USA, Jan. 2019, pp. 83–92

[2] X. Yang, Y. Li, and S. Lyu, (2019) ‘‘Exposing deep fakes using inconsistent head poses,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 8261–8265

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, (2017) ‘‘Densely connected convolutional networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2261–2269

[4] M. S. Rana and A. H. Sung, (2020) ‘‘DeepfakeStack: A deep ensemble-based learning technique for deepfake detection,’’ in Proc. 7th IEEE Int. Conf. Cyber Secure. Cloud Comput. (CSCloud)/6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom), New York, NY, USA, Aug. 2020, pp. 70–75

[5] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, (2020) ‘‘On the detection of digital face manipulation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 5780–5789

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, (2018) “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation”. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[7] D. G¨uera and E. J. Delp, (2018) “Deepfake Video Detection Using Recurrent Neural Networks”. In IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS), 2018.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, (2017) “Two-Stream Neural Networks for Tampered Face Detection”. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1831–1839, July 2017.

[9] Y. Li, M. Chang, and S. Lyu, (2018) “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, pp. 1–7, December 2018.

[10] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, (2019) “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,” Workshop on Applications of Computer Vision and Pattern Recognition to Media Forensics with CVPR, pp. 80–87, 2019.

[11] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, (2019) “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos,” arXiv:1906.06876, June 2019.

[12] G. Patrini, F. Cavalli, and H. Ajder, (2019) “The state of Deepfakes: reality under attack,” Annual Report v.2.3, January 2019.

[13] D. Guera, and E. J. Delp, (2018) “Deepfake Video Detection Using Recurrent Neural Networks,” 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, pp. 1–6, November 2018.

[14] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, (2019) “FaceForensics++: Learning to Detect Manipulated Facial Images,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, pp. 1–11, October-November 2019.

[15] N. T. Do, I. S. Na, and S. H. Kim, (2018) “DeepFakes: Forensics Face Detection from GANs Using Convolutional Neural Network,” International Symposium on Information Technology Convergence (ISITC 2018), South Korea 2018.

[16] H. H. Nguyen, J. Yamagishi, and I. Echizen, (2019) “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, United Kingdom, pp. 2307–2311.

e-ISSN No: 2584-2854

OPEN ACCESS

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED ENGINEERING AND MANAGEMENT (IRJAEM)

Email: editor.irjaem@goldncloudpublications.com
Available online at: https://goldncloudpublications.com/index.php/irjaem

CERTIFICATE OF PUBLICATION

IRJAEM Is Hereby Awarding This Certificate To

Vishal Gawali

In Recognition of The Publication of The Manuscript Entitled

Study on AI Generated Fake-Media Detection

Published In Volume 02 Issue 10 October 2024.

Dr. M. Subramanian
Managing Editor, IRJAEM,
Coimbatore, India

e-ISSN No: 2584-2854

OPEN ACCESS

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED ENGINEERING AND MANAGEMENT (IRJAEM)

Email: editor.irjaem@goldncloudpublications.com
Available online at: https://goldncloudpublications.com/index.php/irjaem

CERTIFICATE OF PUBLICATION

IRJAEM Is Hereby Awarding This Certificate To

Mahesh Gaikwad

In Recognition of The Publication of The Manuscript Entitled

Study on AI Generated Fake-Media Detection

Published In Volume 02 Issue 10 October 2024.

Dr. M. Subramanian
Managing Editor, IRJAEM,
Coimbatore, India

e-ISSN No: 2584-2854

OPEN ACCESS

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED ENGINEERING AND MANAGEMENT (IRJAEM)

Email: editor.irjaem@goldncloudpublications.com
Available online at: https://goldncloudpublications.com/index.php/irjaem

CERTIFICATE OF PUBLICATION

IRJAEM Is Hereby Awarding This Certificate To

Chaturdhan Chaubey

In Recognition of The Publication of The Manuscript

Entitled

Study on AI Generated Fake-Media Detection

Published In Volume 02 Issue 10 October 2024.

Dr. M. Subramanian
Managing Editor, IRJAEM,
Coimbatore, India

e-ISSN No: 2584-2854

OPEN ACCESS

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED ENGINEERING AND MANAGEMENT (IRJAEM)

Email: editor.irjaem@goldncloudpublications.com
Available online at: https://goldncloudpublications.com/index.php/irjaem

CERTIFICATE OF PUBLICATION

IRJAEM Is Hereby Awarding This Certificate To

Akash Gidde

In Recognition of The Publication of The Manuscript Entitled

Study on AI Generated Fake-Media Detection

Published In Volume 02 Issue 10 October 2024.

Dr. M. Subramanian
Managing Editor, IRJAEM,
Coimbatore, India

e-ISSN No: 2584-2854

OPEN ACCESS

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED ENGINEERING AND MANAGEMENT (IRJAEM)

Email: editor.irjaem@goldncloudpublications.com
Available online at: https://goldncloudpublications.com/index.php/irjaem

CERTIFICATE OF PUBLICATION

IRJAEM Is Hereby Awarding This Certificate To

Nilesh Bhelkar

In Recognition of The Publication of The Manuscript Entitled

Study on AI Generated Fake-Media Detection

Published In Volume 02 Issue 10 October 2024.

Dr. M. Subramanian
Managing Editor, IRJAEM,
Coimbatore, India

IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

AI-Fabricated Image Detection

[1] Vishal Ganesh Gawali, [2]Chaturdhan Chaubey, [3]Mahesh Gaikwad, [4]Akash Gidde, [5]Nilesh Bhelkar

[1]Bachelor Of Engineering, [2]Bachelor Of Engineering, [3]Bachelor Of Engineering, [4]Bachelor Of Engineering, [5]Assistant Professor.

[1]Artifficial Intelligence and Data Science,

[1]MCT`s Rajiv Gandhi Institute of Technology, Mumbai, India

Abstract: Deepfake technology, which evolved, transformed digital media at a time when there also arose significant security issues, some of which pose a threat to the legitimacy of content and of system protection. Deepfakes have the ability to spread misinformation that can lead to identity theft and harmful games, and therefore it is necessary to detect them. At first glance, this paper shows how an advanced deepfake detection is built based on Inception and MobileNet base learners along with a CNN based meta learner for the classification decision. An accurate computational speed is achieved in the designed system as well as in the accuracy improvement. Our method has the detection accuracy improvements over current industry standards. Deepfake technology leads to the security and the media verification challenges, which are also discussed in the paper, and the necessity of vendor responsible protection measures is on the foreground.

Keywords— *Deepfake, DeepfakeStack, GANs, Deep Ensemble Learning, Machine learning.*

I. INTRODUCTION

Today, authenticating content from AI generated content poses a new challenge for digital media authentication. The network structure behind deepfake technology of this type powered by GAN is to create highly realistic artificial, videos and images. The research advancements brought a number of advantages to entertainment and creative fields, but at the same time, these can be used as a tool to make the world the simple areas full of misinformation and cyber theft and fraud.CNN based and attention mechanism and ensemble learning model frameworks have been used developed for multiple detection approaches for deepfake content by the researcher.

In the research, ensemble learning techniques are implemented to develop improved deepfake detection methods. On the basis of InceptionV3 and MobileNet as base learners, a CNN based meta learner is proposed to boost the classification results. It achieves interpretability through combination with feature extraction and explainability mechanisms such as Grad CAM, which results in explanations within the extracted features. However, this method shows these experimental tests that it can use to identify the genuine and synthetic media very efficiently, and thus it supports the research in the area of cybersecurity and development in digital forensics.

II. OVERVIEW

A. Definition of Deepfake Detection

Professor Schott focuses his research on the different methods to detect deepfake content using Artificial Intelligence based technology. Deep learning based GANs and autoencoders provide deepfake content so realistic that we cannot reliably distinguish between genuine media and the fake content. These threats against media credibility as well as cybersecurity vulnerabilities and falling public trust are huge, and these are called deepfakes. Machine learning models implemented in automation can be used to find manipulated content from authentic material with high degree of accuracy in differentiation.

B. Project Pipeline

Our project needs to do a multiple stage operation in its workflow for the effective detection of deepfake images. These distinct operational phases are comprised of the detection system.

- Data Collection & Preprocessing
  Instead of building our own dataset we use several public deepfake datasets.
  The MTCNN algorithm is applied for extracting and positioning faces with uniformity for processing.
  All images on the platform get transformed to the square dimension of 256x256 pixels.
- Base Learner Models
  The convolutional structures of InceptionV3 enable the feature extraction at high levels of the spatial space.
  MobileNetV2 offers both good spatial feature extraction functionality and a lightweight set of design that is well suited for live use.

- Feature Aggregation & Meta-Learner
  Outputs of InceptionV3 and MobileNetV2 are merged together and form one structure.
  It also runs final classification through the received features with the help of the CNN-based meta-learner.
- Model Evaluation & Optimization
  Accuracy is along with precision and recall which is coupled up with F1-score and AUC.
  The Grad-CAM has the interpretability feature as it visualizes the regions affected by the manipulations.
- Deployment & Real-Time Detection
  The system enables the users to interact with it through Gradio based UI to send their images for getting instant response from the integrated model.

C. Techniques Used

The approach applies advanced deep learning techniques for making deep fake detection that is accurate and robust.

- Convolutional Neural Networks (CNNs): Used for both feature extraction (base learners) and classification (meta-learner).
- Ensemble Learning: Used by the system to combine intricacy provided by InceptionV3 and MobileNetV2 predictions to build a better accuracy of the system.
- Meta-learner: To achieve stable training operations when Batch Normalization and Dropout are used to minimize overfitting.
- Activation Functions:
  - SoftMax for InceptionV3 (multi-class classification).
  - Sigmoid for MobileNetV2 (binary classification).
  - ReLU is used as activation function for the hidden layers within the meta-learner.
- Grad-Cam Explanation: Users are able to see visual heatmaps that both explain how the model decided and highlight the manipulated regions in deepfake images with Grad-CAM Explainability.

D. Deep Ensemble Learning Technique

The idea behind the ensemble learning approach is to employ multiple base-learners in order to construct an accurate meta learner for the objective of deepfake identification. While making predictions, each individual base-learner designs features on his own before an overall collective decision of the class labels is taken. There are basically two broadly defined ensemble techniques which includes:

a. Using the stacking ensemble model the base models include InceptionV3 and MobileNetV2 are used to give their predictions for training a new CNN-based meta-learner for better final predicted results. Therefore, the use of different representations of the features yields enhanced accuracy results from this approach.
b. FLF has feature combination of deep features as it joins the outcome of base models with the intention of generating new features for the modeling. Finally, a CNN based classifier taking the combined feature representation as input, generating the final prediction with the help of enriched input and the effect of pre avoided overfitting.

Therefore, on the basis of feature extraction and other deep learning layers, combined with ensemble learning approaches, we have an improved detection accuracy without being too much resistant to adversarial attack and being more explainable than the others.

III. RELETEAD WORKS

The research [1] has brought tremendous improvements in the digital face manipulation detection handling of face forgery as an area of study still has a long way to go. Research from 2019 showed modeling these unreliable head poses in deepfakes can enable models to obtain high AUROC scores, suggesting that deepfakes are indeed detectable with near perfect classification accuracy. In fact, these models were particularly targeting the relative difference with respect to the central-face-estimated head pose from full-head and were earmarked as a set of key discriminators for the A. Base Lerner Techniques modified images. However, the study noted major drawbacks concerning the current approach in dealing specifically with those for low quality images which most current techniques cannot adequately handle.

Strengthening adaptability and specificity were the concepts on which [2] builds during 2020, introducing into their detection method an attention mechanism. In cases like this and which remember as this problem actually pertains, being able to concurrently respond with further enhanced detection localization of altered features of facial, especially within these scenarios characterized with reduced rates in false detections, this would stand as better applied. This, however also necessitated greater datasets that would encompass a much broader sweep of manipulation types to more adequately evaluate and improve upon detection methods.

The researcher in [3] pointed out a way forward by improving sensitivity and robustness of detection schemes against the sophisticated and diverse manipulations of faces-in the-wild in the digital era at the same time calling for new methods from the research community to fill the previously identified gaps. Most of the works have been in the identification difficulty of deepfakes and effort to decrease the face manipulations by the observation of the artifact in the images. Their methodology was based on machine learning technique, specifically using classification algorithms of k-NN and logistic regression for image segmentation and analysis identified as a deepfake. When tested with GAN generated data, the k-NN classifier performed strongly (AUC: 0.852), indicating that it is good at discriminating between real and fake images. Moreover, logistic regression models combined with other features can also achieve comparable performance to deep models, showing a potential for simple and light deepfake detection the open-source software he used won the end-to-end experiments for face examination. This feature will just record the time point when a malicious ad appears. Does not guarantee to store names of people who have visited any site on daytimes face current online attacks. But even so, there obviously will need to be technological filings for quick recovery. The need for better adjustment of contact points on structures and organs has often been raised in past experience with interventional systems. A general trend less movable joints in medical equipment the researchers put forward the direction of post operative drainage for patients who undergo hepatectomy many works of this kind are published these days, but there is no

cross-reference standard cataloging formal structure to account for them all, so it's tough sledding indeed, where these do not appear for discussion.

In [4] Dense Net – a new and exciting concept in the construction of convolutional networks. DenseNets are known for their dense connectivity pattern within layers, which leads to a decrease in the number of model parameters compared to conventional architectures such as ResNets. This design not only solves the vanishing gradient problem through improving the gradient flow but also improves the parameter efficiency, making DenseNets have high computational efficiency. For example, a DenseNet that has the same computational complexity as a ResNet-50 can outperform a ResNet-101, providing high computational efficiency. The authors noted several questions that may be addressed in future works, including the study of the capacity of DenseNets that was not studied systematically at the time of writing.

The work in research [5] introduces DeepfakeStack, which is a novel ensemble learning technique dedicated to the improvement of the deepfake detection. This is where DeepfakeStack is superior since it applies a number of deep learning models in one step to improve the detection rates. Thus, this ensemble approach takes the best from different models to design a reliable system that can detect deepfake videos with a high precision as confirmed by the detection accuracy of 99.65%.

This not only provides an effective approach for deepfake detection but also presents possible directions for model enhancement and future deep learning research on combating fake media. The details of literature survey are shown in table.

Author & Ref.	Year	Research Paper
Xin Yang, Yuezun Li and Siwei Lyu [2]	2019	Exposing Deepfakes Using Inconsistent Head Poses
Hao Dan FengLiu Joel Stehouwer Xiaoming Liu Anil Jain [5]	2020	On the Detection of Digital Face Manipulation
Falko Matern Christian Riess Marc Stamminger [1]	2019	Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations.
Laurens van der Maaten, Zhuang Liu, Gao Huang [3]	2017	Densely Connected Convolutional Networks
Md. Shohel Rana , Andrew H. Sung [4]	2020	DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection

Table 1. literature survey detail

IV. RESEARCH METHODOLOGY

Generated fake media that integrates three powerful techniques: Base Learner Creation, Stack Generalization, and Ensemble Learning. Our approach systematically develops and integrates multiple machine learning models to ensure accurate and robust detection of synthetic content. The following steps outline the methodology for implementing this model:

A. Dataset Collection and Preprocessing:

The procedure is initiated with data collection and preprocessing, where large-scale data like RealVsFake are used to provide diversity in facial geometry, lighting, and manipulation methods. Preprocessing involves face detection and cropping with the aid of methods like MTCNN or Dlib, normalization and standardization of pixel intensity, and data augmentation methods like rotation, flipping, color alterations, and noise injection. These processes ready the data for training and strengthen the model's power to generalize over varied situations.

B. Base Learner Creation:
The second stage includes base learner training, where several CNN models are trained to identify various features of deepfake images. Models such as ResNet-50, EfficientNet-B3, and XceptionNet are used as base learners, each of which is trained to identify particular patterns like hierarchical spatial features, fine-grained textures, and global inconsistencies. These models are initialized with ImageNet pretrained weights and fine-tuned on deepfake datasets. Following training, feature vectors are derived from the last layers of every CNN, which are combined to form an overall representation of the input data. Feature fusion methods involve concatenation, weighted averaging, and attention mechanisms to ensure that the most important features.

C. Meta Learner Creation:
A meta-learner is brought in to improve predictions by learning from base learner outputs. This secondary CNN is a classifier at the high level, aggregating individual base learner strengths and enhancing overall classification performance. The meta learner avoids overfitting and improves generalization by taking advantage of multiple independent CNN outputs. Techniques such as batch normalization, dropout, and adaptive learning rates are also used to further optimize the model. Data balancing ensures that real and synthetic images are represented equally in training, avoiding bias and enhancing model robustness.

D. Prediction and Evaluation:
During the prediction and evaluation phase, the trained model labels images or videos as real or synthetic with a corresponding confidence score. For better interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to identify highlighted relevant regions within the input data that led to the output prediction. Such an explainability technique enables researchers to comprehend the reasoning process of the model and instill confidence in its predictions. The performance of the model is measured by quantitative metrics including accuracy, precision, recall, F1 score, and AUC-ROC curves, as well as qualitative assessment through visual examination of Grad-CAM heatmaps.

E. Prediction and Evaluation:
the approach finishes with future research focused on improving deepfake detection further. This involves the addition of Vision Transformers (ViTs) for enhanced feature representation, adversarial training to defend against changing deepfake methods, and investigating multi-modal detection through integrating visual and audio signals. With the use of ensemble learning, feature fusion, and explainability methods, this approach offers a solid framework for detecting AI synthesized fake media with high accuracy and dependability in practical applications.

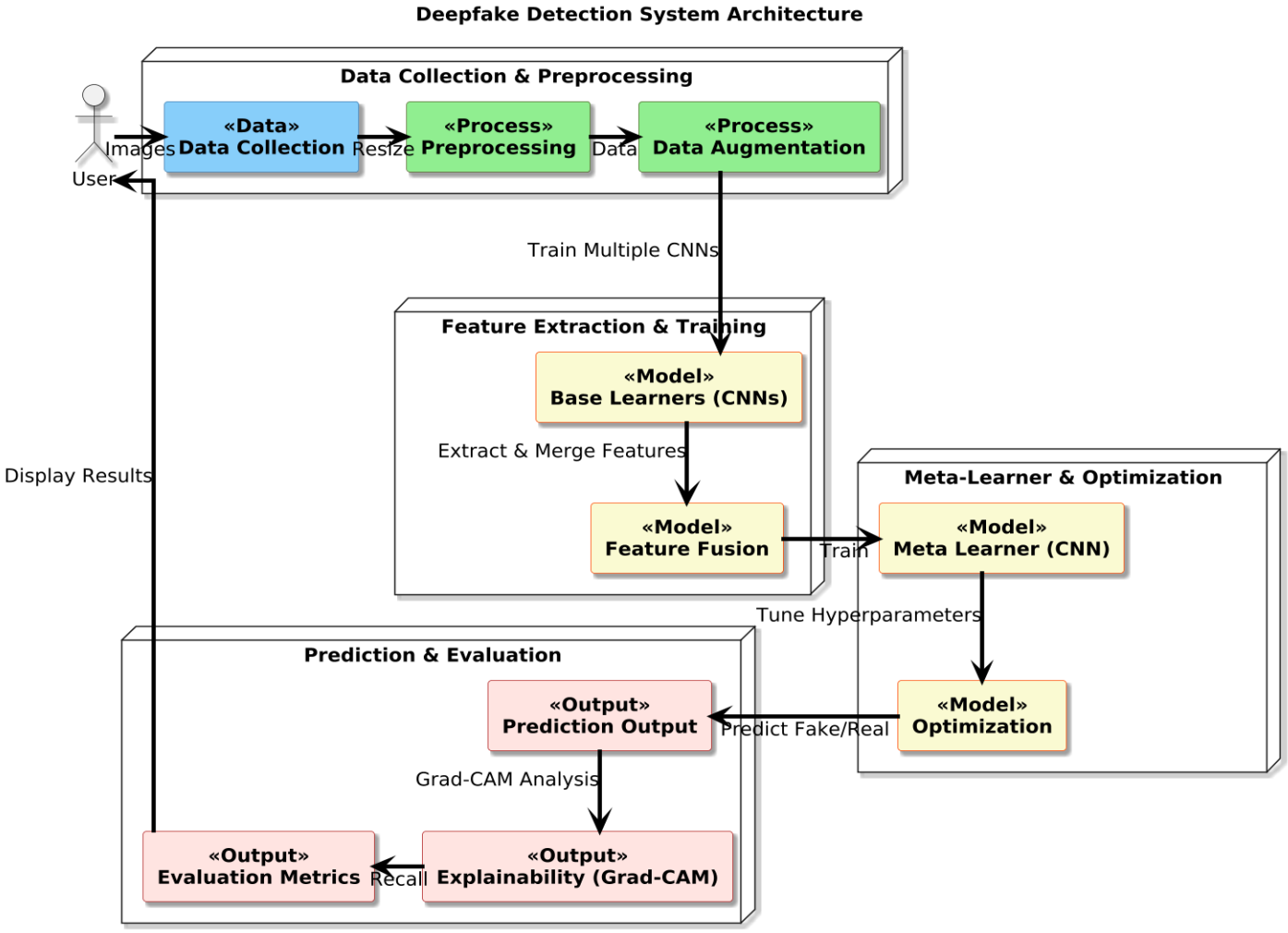


Fig. 1 System Architecture

V. RESULTS AND DISCUSSION

The deepfake detection framework proposed, using CNN-based ensemble learning, was comprehensively tested to determine its performance, resilience, and explainability. The findings show the efficacy of the approach in differentiating between genuine and AI-created fake media. A comparative analysis of MobileNet, InceptionNet, and the Ensemble-based CNN method was carried out to emphasize their respective strengths and limitations. The following is a critical analysis of the findings, emphasizing quantitative measures, qualitative observations, and comparisons with current best practices.

A. Precision

Precision is the ratio of true positive instances to all positive instances predicted.

In this research:

- Inception and the Ensemble model had 100% precision, meaning that all positive predictions by these models were accurate.
- Mobilenet had 96% precision, with a slightly higher chance of false positives than the other models.

B. Recall

Recall, or sensitivity, is the rate of actual positive instances correctly identified by the model.

The output indicates:

- The Inception and Ensemble model recorded 100% recall, as they were capable of detecting all positive instances without leaving any out.
- Mobile net recorded 97% recall, which signifies a slight trade off in detecting all positive instances

C. F1-Score

The F1 score is the harmonic mean of recall and precision, giving a balanced measure of a model's performance.

The findings show:

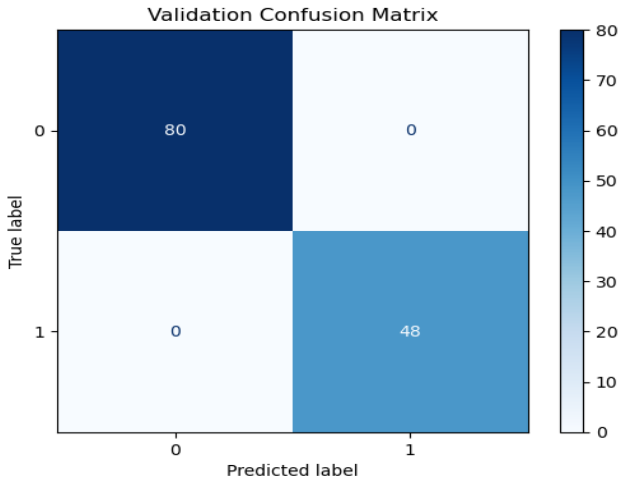
- Inception and the Ensemble model had 100% F1 score, indicating their ideal balance between recall and precision.
- Mobilenet had 97% F1 score, indicating strong but slightly poorer performance than the other models.

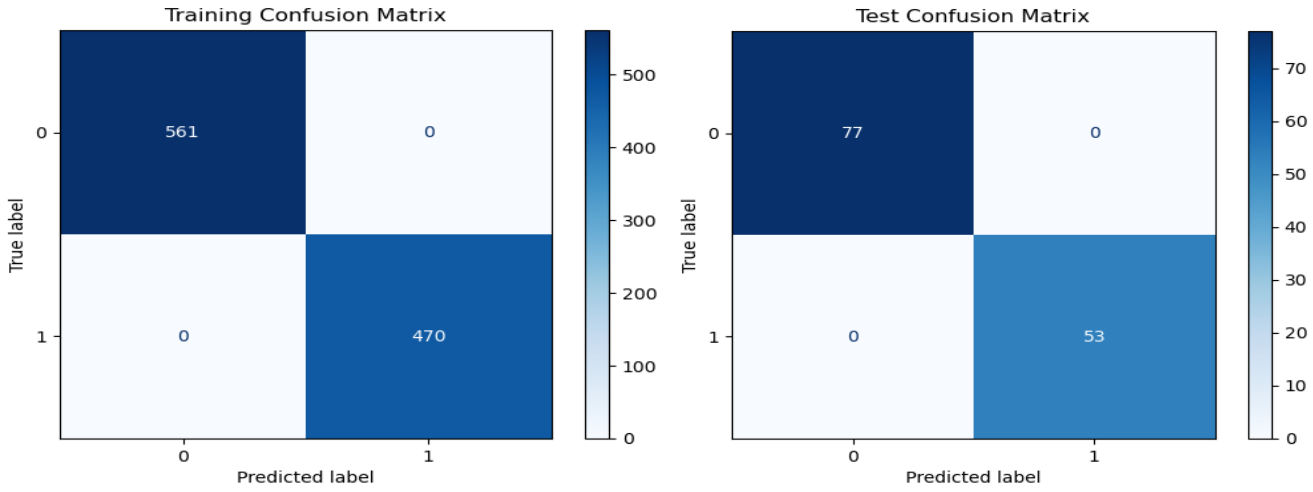
D. Test-Accuracy

Test accuracy quantifies the overall accuracy of the model's predictions on new data. The Ensemble model's 100% performance on precision, recall, and F1 score indicates that it is most likely to have the highest accuracy. Mobile net, with slightly lower scores, would probably have a slightly lower accuracy.

Comparative Study of Models

Metrix	Inception	Mobilenet	Ensemble
Precision	100%	98%	100%
Recall	100%	99%	100%
F1 Score	100%	98%	100%
Test Accuracy	100%	99%	100%





In order to further assess the performance of the suggested deepfake detection model, confusion matrices were created for the training set, validation set, and test set. The matrices give a detailed description of the predictions of the model, including true positives, true negatives, false positives, and false negatives.

REFERENCES

[1] F. Matern, C. Riess, and M. Stamminger, (2019) ‘‘Exploiting visual artifacts to expose deepfakes and face manipulations,’’ in Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW), Waikoloa Village, HI, USA, Jan. 2019, pp. 83–92

[2] X. Yang, Y. Li, and S. Lyu, (2019) ‘‘Exposing deep fakes using inconsistent head poses,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 8261–8265.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, (2017) ‘‘Densely connected convolutional networks,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2261–2269

[4] M. S. Rana and A. H. Sung, (2020) ‘‘DeepfakeStack: A deep ensemble-based learning technique for deepfake detection,’’ in Proc. 7th IEEE Int. Conf. Cyber Secure. Cloud Comput. (CSCloud)/6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom), New York, NY, USA, Aug. 2020, pp. 70–75

[5] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, (2020) ‘‘On the detection of digital face manipulation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 5780–5789. Adversarial Networks for Multi-Domain Image-to Image Translation”. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, (2018) “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to Image Translation”. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[7] D. G¨uera and E. J. Delp, (2018) “Deepfake Video Detection Using Recurrent Neural Networks”. In IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS), 2018.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, (2017) “Two-Stream Neural Networks for Tampered Face Detection”. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1831–1839, July 2017.

[9] Y. Li, M. Chang, and S. Lyu, (2018) “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, pp. 1–7, December 2018.

[10] I. E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, Masi, and P. Natarajan, (2019) “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,” Workshop on Applications of Computer Vision and Pattern Recognition to Media Forensics with CVPR, pp. 80–87, 2019.

[11] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, (2019) “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos,” arXiv:1906.06876, June 2019.

[12] G. Patrini, F. Cavalli, and H. Ajder, (2019) “The state of Deepfakes: reality under attack,” Annual Report v.2.3, January 2019.

[13] D. Guera, and E. J. Delp, (2018) “Deepfake Video Detection Using Recurrent Neural Networks,” 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, pp. 1–6, November 2018.

[14] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, (2019) “FaceForensics++: Learning to Detect Manipulated Facial Images,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, pp. 1–11, October-November 2019.

[15] N. T. Do, I. S. Na, and S. H. Kim, (2018) “DeepFakes: Forensics Face Detection from GANs Using Convolutional Neural Network,” International Symposium on Information Technology Convergence (ISITC 2018), South Korea 2018.

Certificate of Publication

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-ISSN: 2348-1269, P-ISSN: 2349-5138

The Board of

International Journal of Research and Analytical Reviews (IJRAR)

Is hereby awarding this certificate to

Vishal Ganesh Gawali

In recognition of the publication of the paper entitled

AI-Fabricated Image Detection

Published In IJRAR (www.ijrar.org) UGC Approved - Journal No : 43602 & 7.17 Impact Factor

Volume 12 Issue 2 April 2025, Date of Publication: 04-April-2025

PAPER ID : IJRAR25B1116

Registration ID : 309689

R.B.Joshi

EDITOR IN CHIEF

UGC and ISSN Approved - Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS | IJRAR

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijrar.org | Email: editor@ijrar.org | ESTD: 2014

Manage By: IJPUBLICATION Website: www.ijrar.org | Email ID: editor@ijrar.org

IJRAR | E-ISSN: 2348-1269, P-ISSN: 2349-5138

Certificate of Publication

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-ISSN: 2348-1269, P-ISSN: 2349-5138

The Board of

International Journal of Research and Analytical Reviews (IJRAR)

Is hereby awarding this certificate to

Chaturdhan Chaubey

In recognition of the publication of the paper entitled

AI-Fabricated Image Detection

Published In IJRAR (www.ijrar.org) UGC Approved (Journal No : 43602) & 7.17 Impact Factor

Volume 12 Issue 2 April 2025, Date of Publication: 04-April-2025

PAPER ID : IJRAR25B1116

Registration ID : 309689

R.B.Joshi

EDITOR IN CHIEF

UGC and ISSN Approved - Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS | IJRAR

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijrar.org | Email: editor@ijrar.org | ESTD: 2014

Manage By: IJPUBLICATION Website: www.ijrar.org | Email ID: editor@ijrar.org

IJRAR | E-ISSN: 2348-1269, P-ISSN: 2349-5138

Certificate of Publication

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-ISSN: 2348-1269, P-ISSN: 2349-5138

The Board of

International Journal of Research and Analytical Reviews (IJRAR)

Is hereby awarding this certificate to

Mahesh Gaikwad

In recognition of the publication of the paper entitled

AI-Fabricated Image Detection

Published In IJRAR (www.ijrar.org) UGC Approved (Journal No : 43602) & 7.17 Impact Factor

Volume 12 Issue 2 April 2025, Date of Publication: 04-April-2025

PAPER ID : IJRAR25B1116

Registration ID : 309689

R.B.Joshi

EDITOR IN CHIEF

UGC and ISSN Approved - Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS | IJRAR

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijrar.org | Email: editor@ijrar.org | ESTD: 2014

Manage By: IJPUBLICATION Website: www.ijrar.org | Email ID: editor@ijrar.org

IJRAR | E-ISSN: 2348-1269, P-ISSN: 2349-5138

Certificate of Publication

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-ISSN: 2348-1269, P-ISSN: 2349-5138

The Board of

International Journal of Research and Analytical Reviews (IJRAR)

Is hereby awarding this certificate to

Akash Gidde

In recognition of the publication of the paper entitled

AI-Fabricated Image Detection

Published In IJRAR (www.ijrar.org) UGC Approved (Journal No : 43602) & 7.17 Impact Factor

Volume 12 Issue 2 April 2025, Date of Publication: 04-April-2025

PAPER ID : IJRAR25B1116

Registration ID : 309689

R.B.Joshi

EDITOR IN CHIEF

UGC and ISSN Approved - Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS | IJRAR

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijrar.org | Email: editor@ijrar.org | ESTD: 2014

Manage By: IJPUBLICATION Website: www.ijrar.org | Email ID: editor@ijrar.org

IJRAR | E-ISSN: 2348-1269, P-ISSN: 2349-5138

Certificate of Publication

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-ISSN: 2348-1269, P-ISSN: 2349-5138

The Board of

International Journal of Research and Analytical Reviews (IJRAR)

Is hereby awarding this certificate to

Nilesh Bhelkar

In recognition of the publication of the paper entitled

AI-Fabricated Image Detection

Published In IJRAR (www.ijrar.org) UGC Approved (Journal No : 43602) & 7.17 Impact Factor

Volume 12 Issue 2 April 2025, Date of Publication: 04-April-2025

PAPER ID : IJRAR25B1116

Registration ID : 309689

R.B.Joshi

EDITOR IN CHIEF

UGC and ISSN Approved - Scholarly open access journals, Peer-reviewed, and Refereed Journals, Impact factor 7.17 (Calculate by google scholar and Semantic Scholar | AI-Powered Research Tool) , Multidisciplinary, Monthly Journal

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS | IJRAR

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijrar.org | Email: editor@ijrar.org | ESTD: 2014

Manage By: IJPUBLICATION Website: www.ijrar.org | Email ID: editor@ijrar.org

IJRAR | E-ISSN: 2348-1269, P-ISSN: 2349-5138